知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

数据出版的实践模式对比研究
——以地球科学领域为例

◎韩露[1] 丁毅[2]
[1] 北京理工大学图书馆 北京 100081
[2] 中国地质科学院地质研究所 北京 100037

摘要：[目的/意义]科学数据出版是数据密集型科学发现的主要学术传播方式，对于实现数据重用、科学验证具有重要的意义。[方法/过程]地球科学从原有的数据共享模式到目前的数据出版发生了巨大的变化。作者将数据出版分为数据期刊出版、数据仓储出版、数据和论文联合出版 3 种模式，对于每一种模式的实践方法和关键要素进行统计和对比，重点分析三种模式的优劣、同行数据评议问题和地学数据出版中分层元数据的重要性。[结果/结论]通过研究，作者提出基于仓储的出版便于融入科学数据管理过程，有利于数据重用，但是此类出版方式缺少同行评议；数据的同行评议应该有别于学术论文，注重数据在参与科研和产生再生性成果的过程中的重用性；元数据的分层描述对于地学大数据的保存和重用都具有重要意义。

关键词：数据出版 数据仓储 数据期刊 地学数据共享
分类号：G237.6

引用格式：韩露，丁毅. 数据出版的实践模式对比研究——以地球科学领域为例 [J/OL]. 知识管理论坛，2019, 4(3): 152-162[引用日期]. http://www.kmf.ac.cn/p/173/.

地球科学（以下简称“地学”）是一个数据科学，但是由于数据采集难度大、空间范围广、仪器设备价值昂贵等问题，数据共享、获取和重用一直都是地学研究的重要内容。20 世纪早期，人们采用穿孔卡片的方式来记录数据。20 世纪 70 年代，为实现地学数据的共享，多个国际组织先后成立，如世界数据中心（World Data Center，简称 WDC，2008 年后被 World Data System 简称 WDS 取代）、地球观测组织（Group on Earth Observations, GEO）、地球观测数据网（Data Observation Network for Earth）。1988 年中国加入 WDS 并成立了 9 个数据中心，多数为地球科学领域的数据中心，如地震、地质、地球物理数据中心[1]。2002 年度科学数据共享工程启动了“地球系统科学数据共享网项目”，建立

基金项目：本文系中国地质科学院基本科研业务项目“深部探测知识库的建设与研究”（项目编号：JB1602）研究成果之一。
作者简介：韩露（ORCID：0000-0002-1210-7388），副研究馆员，博士；丁毅（ORCID：0000-0002-7952-5867），副高级工程师，博士，通讯作者，E-mail：dingyi@cags.ac.cn。
收稿日期：2019-03-04 发表日期：2019-06-14 本文责任编辑：刘远颖

知识管理论坛，2019(3):152–162
DOI：10.13266/j.issn.2095–5472.2019.016

了地球科学数据共享平台[2]。但是这些数据中心和共享平台的数据获取往往受到限制，共享不活跃，甚至形成了数据孤岛。褚云强等[3]对科学大数据的共享机制进行了研究，提出缺乏数据共享的政策和机制是阻碍其发展的主要原因，而调动科研人员主动共享数据积极性的重要机制就是自下而上的科学数据出版。数据出版是近几年由出版界和数据共享界共同提出的新概念[4]，在开放科学的趋势下，一些国家的资助机构和国际组织纷纷出台政策要求管理和共享研究数据。如美国 NSF 地学学部要求将完整的数据集、数据产品、软件和集成整合数据必须在两年内可公开访问[5]。由于数据中心要求用户在使用数据时体现数据作者的贡献，学术期刊要求作者在发表文章时提供与科学结论相关的数据，这两种机制最终成为推动数据出版的共同动力。数据出版的核心是为数据引用提供标准的引用格式和永久访问地址，使科学数据是可获取、可理解、可评估、可使用的[6]，与原有的数据共享相比，数据出版更加强调了数据重用的可能性以及数据被科研人员的认可程度。德国是首个研究科学数据出版的国家，德国国家科学技术图书馆申请成为国际上首个科学数据 DOI 代理机构，并为很多原 WDS 数据中心的数据资源进行了 DOI 注册[7]，这些数据中心逐步开展数据出版实践，数据出版已经成为共享的新的形式。数据引用评价体系的建立，客观反映了数据贡献者的影响力，改善了原有的数据共享不活跃的状况。我国的地学数据出版还在起步阶段，与原有的数据共享形成了互补的态势，中国地质调查局已经开展了数据的 DOI 注册。地学领域的数据出版由于数据共享起步较早，在数据出版实践中具有代表性，一些传统的地学期刊论文很早就有将数据作为附件同时出版的先例，如《岩石学报》。笔者通过调研地学领域的各类数据出版模式的实践方法，探讨在地球科学领域数据出版中存在和亟待解决的主要问题。

1 地球科学领域数据出版的主要模式

自从数据出版的概念被提出后，一些 WDS 逐渐将原有的数据共享服务转化为数据的存储管理和出版服务，注重数据的发现、获取、重用、回溯等。笔者主要检索了 re3data.org 平台中注册的数据中心仓储，其中地球科学领域已经注册的仓储有 647 个，仅次于生物学领域的数据仓储数量，是数据出版实践为数较多的学科。笔者利用 JCR（Journal Citation Reports），对地球科学下所包含的地质、地球物理、地球化学、环境科学、地理、地球科学与多学科交叉的期刊进行了查询，同时参照科睿唯安的 DCI（Data Citation Index）中地球科学领域的数据条目的出版源和期刊网站的介绍进行了识别。JCR 收录的地球科学类数据期刊主要有三种：*GeoScience Data Journal*（简称 GDJ）、*Scientific Data*（简称 SD）、*Earth System Science Data*（简称 ESSD）。我国于 2014 年发行了地球科学领域的数据出版平台全球变化科学研究数据出版系统（Global Change Research Data Publishing & Repository，GCRDPR），2017 年该出版系统推出了《全球变化数据学报》。一些数据仓储和声望较好的学术期刊进行合作，以期增加数据的影响力，实现学术论文和数据的互联。在地球科学领域的数据仓储中 Pangaea 出版的数据量较多，Elsevier 的 Science Direct 和 Scopus 均与 Pangaea 仓储进行了联合，实现学术论文和数据的集成出版。还有一类是将数据作为附件与学术论文一起出版，但是这类出版的主体依然是学术论文，本文不做讨论。综上所述，地学领域主要呈现的数据出版模式包括：

（1）数据期刊：将数据转化为论文的形式进行描述，发表在数据期刊中，通常与领域内的数据中心或公共数据仓储联合实现数据集的存缴，数据论文的出版具有和学术论文相似的同行评议评审流程。

（2）基于仓储的直接数据出版：通过数据仓储发布数据，部分由原来的数据共享平台转

变而来，数据出版作为数据管理和共享的一种方式。

（3）学术论文和数据的联合出版：数据和学术论文之间的互联可实现数据和科学发现的精确关联和验证。与上述两种数据出版本质的区别是出版的内容不仅仅是数据，还包括来自于数据的科学性发现等学术论文。目前，大多建立在出版社与数据仓储平台合作的基础上，通过增强出版的形式实现论文和数据的关联性出版与发布。

2 地学数据期刊

数据期刊与传统的科研传播中的学术论文的出版重点不同，主要是提供和描述数据集本身，通常不关注结论的科学创新性与否，出版的主要目的是提高数据的影响力，使科学数据能够更多地被描述、解释、重用。笔者对比了 4 种地球科学领域的数据期刊（见表 1），各期刊的载文量统计见图 1，其中 SD 只统计了地球与环境科学主题下的载文量。

表 1 地球科学类数据期刊调查表

数据期刊	GDJ	SD	ESSD	GCRDPR
创刊年	2014	2014	2009	2014
出版者	Wiley Online Library	Nature	Copernicus Puli-cations	《全球变化数据学报（中英文）》编辑部
主题范围	地球科学领域的跨学科期刊	综合性期刊，包含地球科学与环境科学主题	地球科学领域的跨学科期刊	地理科学
期刊类型	专业型	综合型	专业型	专业型
同行评议流程	期刊内部的同行评议流程	期刊内部的同行评议流程	网络开放的两段式同行评议方法	内部专家匿名评议
JCR 影响因子（2016）	2.800	4.836	6.696	无
数据集的访问方法	提交到建议的仓储中	提交到建议的数据仓储中	提交到建议的仓储中	期刊系统提供的数据仓储
数据集引用标准	DataCite	自定义标准	DataCite	自定义标准
出版方式	在线	在线	在线和纸行	在线

2.1 保证数据的可访问性

在保证数据的可访问性方面，所有的期刊都需要将数据提交到数据仓储中进行长期保存，并提供可持续引用的唯一标识符，目前大多数仓储都采用了 DOI 作为地址解析代理的标准。GDJ、SD、ESSD 自身都没有保存数据的仓储，因此刊物列出了可提交数据的仓储列表，在论文提交的过程中要同时选择一个建议的仓储保存数据。大部分数据期刊都是利用 re-3data.org 平台中注册的仓储来选取建议提交的仓储，或者选取与数据期刊内容相符的国家级数据中心，这种出版模式需要出版机构和数据保存管理机构的合作，对数据的管理和数据论文的出版是由两个独立的系统协作完成的。但是 GCRDPR 与其他 3 种期刊略有区别，它本身是一个网络出版的系统，集元数据、实体数据、数据论文的出版于一体。首次提交数据后该出版平台会根据元数据的质量和数据描述判断是否可以接受该数据，如果接受再通知作者参照指南撰写数据论文并分配 DOI 给数据集，然后进入数据论文同行评议的流程。

2.2 数据论文的写作规范

数据论文是对数据集内容的增强性描述，不同学科领域对于数据论文的描述维度各有差异，笔者将 4 种地学数据期刊对数据集描述的内容框架进行了汇总，共涉及 9 个方面：

知识管理论坛，2019(3):152–162
DOI：10.13266/j.issn.2095–5472.2019.016

①数据集的有效访问方式，比如提供 DOI 或者 URI；②数据集的覆盖范围，包括时间和空间的覆盖范围；③数据集的格式信息，如数据本身的格式、编码方法和编码语言等；④数据集的授权许可，用于支持数据集的合法使用；⑤项目信息，提供数据集的生产信息，如生产数据集的目的和资助来源；⑥来源信息，提供生产数据集的方法的描述信息，包括采用的工具、处理方法；⑦质量信息，提供数据质量的描述信息，如数据集的局限性和异常信息；⑧重用信息，提供数据集的使用方法描述，促进数据集重用的可能；⑨对于数据引用的支持与规范。表 2 对比了 4 种期刊各自的内容框架。ESSD 没有明确规定其描述的内容要素，但是在发表数据论文的过程中编辑会建议作者去关注决定数据论文潜在价值的相关内容，给出评议的指南，指南中包含的内容与表 2 中列出的 9 个方面基本相同。《全球变化科据学报》没有对数据论文的内容做统一的规范，而是在元数据中进行了规范。

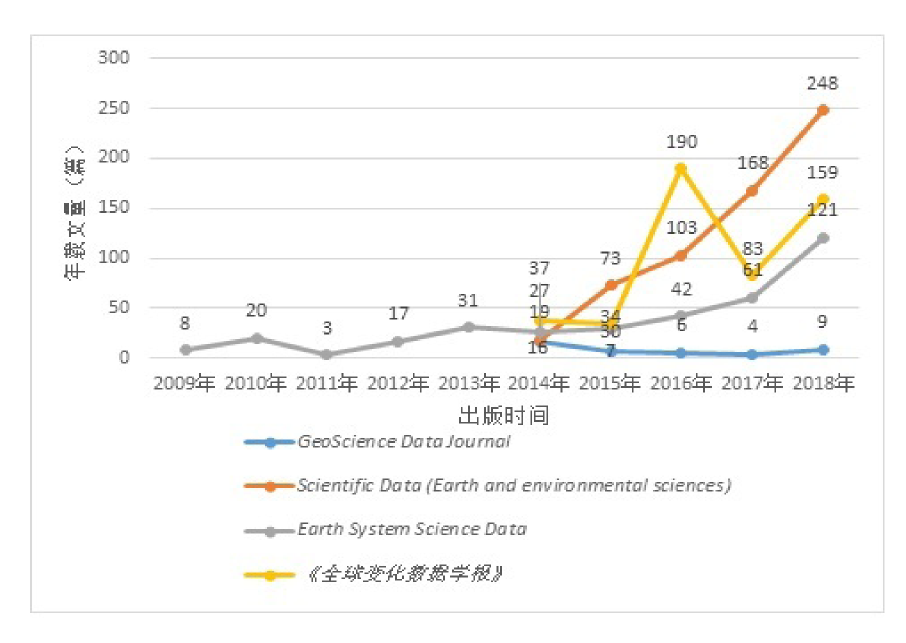


图 1 各期刊的载文量统计

表 2 4 类期刊对数据集描述要求的对比

期刊名称	GDJ	SD	ESSD	GCRDPR
数据集可获取性标识（DOI 或 URI）	√	√	√	√
数据集的覆盖范围	√			√
数据集格式信息	√	√		√
数据集授权许可信息		√		
项目信息				√
数据重用信息	√	√		√
数据质量描述信息		√		
数据来源信息	√	√		√
对于数据引用的支持与规范	√	√	√	√

2.3 同行评议与质量控制

在同行评议方面，SD、GDJ 和 GCRDPR 采用了与传统学术期刊类似的同行评议流程，但是评议的内容则更注重数据版权、质量。GCRDPR 专门提供了一个评议内容的模板，包含数据集的意义、质量、学术相关性、作者的知识产权意识、数据的可获取性共 5 个方面的内容。ESSD 与其他 3 种期刊略有不同，采用了两段式的开放式同行评议，即作者提交论文初稿后将在网络平台上作为讨论稿出版，然后经过专家的评审后，被专业编辑评估，但不评估其科学内容，仅对数据论文是否符合论文的撰写要求、是否符合期刊的主题范畴进行评估，并提出技术性的修改建议。论文在这个状态将保持 8 周，期间可以进行各种同行评价和讨论等，每篇论文接受至少两个专家的终审。在公开讨论阶段，作者需要回复各种评论并修改论文初稿，最后才能决定是否出版。对于数据论文的同行评议，出版界尚未有统一的标准，在这种情况下这种两段式的公开评议方式更有利于真正的数据使用者或学术同行提出较为准确的意见。

3 基于地学仓储的数据管理和出版服务

3.1 基于仓储直接出版数据的要素

基于仓储实现数据出版是将原有共享数据的方式转变为出版后出现的，目前各个领域没有一个明确标准规定其应具备的要素。J. E. Kratz [8] 在对数据出版的调查中发现大多数研究人员更关注以下几点：数据是否有唯一标识、是否可开放获取、是否存放在一个仓储中、是否具有正式丰富的元数据。自从 re3data.org 提供了对世界各国研究数据长期保存和管理的仓储的注册服务后，仓储的建设得以规范，仓储的影响力也得到了提升，一些出版商和期刊（如 PeerJ、Springer 和 Nature 的 Scientific Data 等）都以 re3data.org 作为简单的工具来识别适合的数据仓储。笔者选取表 1 中的数据期刊推荐的数据仓储进行了对比，发现在数据出版方面具备的共同要素包含以下 5 个方面：

（1）为数据提供唯一标识，保证数据具有一个固有地址可供访问。

（2）提供数据提交编辑工具和进行数据质量控制。数据提交编辑工具包括对数据的提交、对数据说明的编辑、对元数据的编辑。数据质量控制大多数是由数据管理员完成，保证上述提交和编辑的内容的完整性和一致性。

（3）提供正式的数据使用许可声明。仓储不具有像期刊同样的商业版权，数据使用许可在一个开放的科学环境下对数据创建者和使用者提供双重保护。为了避免在数据重用过程中的各类权益纠纷，需要提供一个数据使用许可声明。

（4）提供正式的数据引用格式。数据的正式引用是重用数据实现数据定位的机制 [9]，T. E. Pronk 等 [10] 在博弈论框架下分析了共享和出版数据的影响因素，结果表明与政策规定相比，降低成本和增加引用等更具有激励效果，即正式的引用对于提升科研人员出版的积极性具有促进作用，引用是保障数据作者与管理者数据权益的一种有效方式。

（5）数据仓储可开放获取数据。数据可获取是实现重用的前提，大部分提供数据出版服务的仓储都是可开放获取的，少数仓储需要权限才能获取数据。

作者在 re3data.org 中按照上述 5 个要素进行了检索，通过统计发现在开放获取方面大部分仓储都满足，只有少数是分级开放的。在数据标识符方面，目前有 234 个仓储提供了此服务，采用较多的标识符为 DOI、ARK、URI。其中 101 个仓储采用了 DOI，其他未提供此服务的仓储多数是采用了外部公共仓储的注册服务来实现自身的数据管理。在数据提交编辑与质量控制方面，与期刊不同的是，在仓储中数据质量控制由数据管理人员完成，属于技术性的审查，不同于科学性的同行评议，这种技术审

知识管理论坛，2019(3):152–162
DOI：10.13266/j.issn.2095–5472.2019.016

查的方式主要有两种：①一部分仓储的数据的专业性较强，且与项目密切相关，在这类仓储的数据质量控制流程中增加了与项目相关的同领域专家来验证数据，如美国冰雪数据中心就采用了这种方式；②一部分仓储支持仓储的认证标准并通过了认证，其中支持认证比较多的有 WDS（世界数据中心）[11]、DSA（荷兰的数据归档和网络服务发布的数据批准印章）[12] 和 CoreTrustSeal[13]，其中有 48 个仓储属于 WDS。对于数据使用的声明采用较为广泛的是知识共享家族的 CC0、CC-BY、CC-BY-SA 许可协议 [14]。对于正式的数据引用方面，由于数据出版起步较晚，国内外尚无统一标准，地学数据出版中较为广泛参考的是 RORCE11 数据引用联合声明 [15] 和 DataCite 的出版与引用方案 [16]。而在具体的应用中，一些数据仓储要求在发表学术论文使用数据的同时采用仓储提供的正式标引，还有一些数据仓储要求在论文致谢中声明数据的来源。前者多数为与期刊合作的数据出版服务仓储，出版的数据大多是经过挑选、处理后的成果性数据；后者多用于国家级地学研究机构下的数据中心，如美国的 NOAA（美国国家海洋气象局），USGS（美国地质调查局）等，共享的数据多为大规模的原始采集数据或基础地理数据，共享方式除了数据出版，还包括 FTP 服务或者 RESTAPI 等。上述统计结果见图 2，在 re3data 网站注册的仓储仅有 26% 完全满足上述 5 个要素。

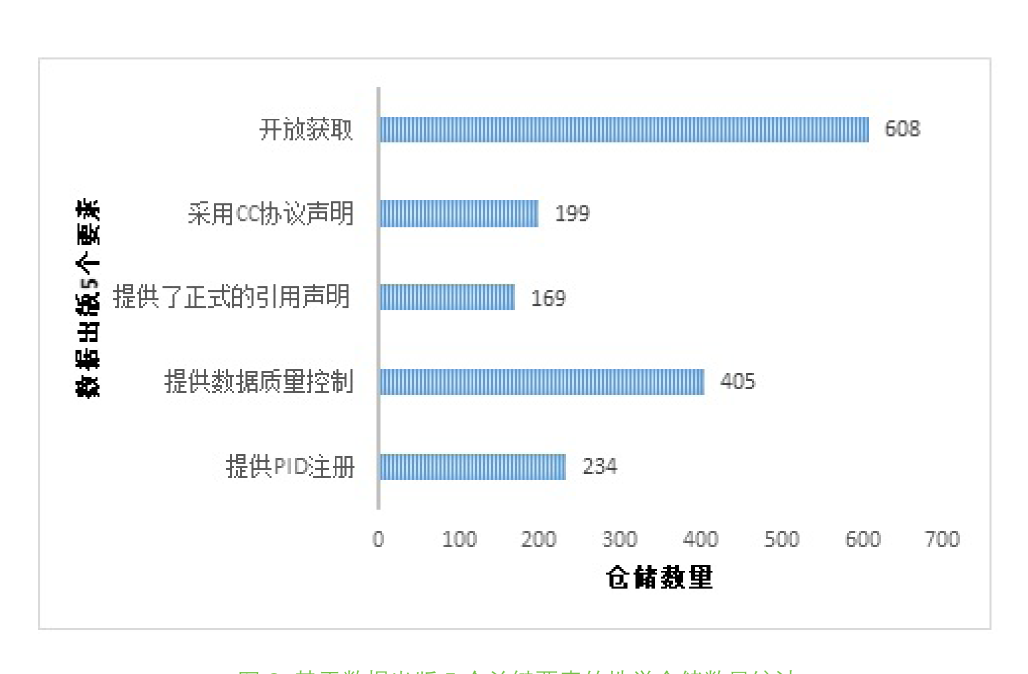


图 2 基于数据出版 5 个关键要素的地学仓储数量统计

3.2 基于科学数据管理过程的对比

数据出版与研究过程中的数据管理息息相关，因此不存在绝对独立的数据出版或数据管理平台 [17]。在 re3data.org 的地学领域数据仓储中，提供数据出版服务较多的仓储主要有以下 3 类：①公共的数据出版平台；②地学领域的国际数据共享组织或数据中心，如 WDS；③国家级的地学研究机构所建设的数据汇交共享平台。笔者对数据管理过程中与出版相关的内容进行了归纳，选取具有代表性的数据仓储进行了对比，如表 3 所示：



知识管理论坛
2019 年 第 3 期 (总第 21 期)

表 3 三类仓储的数据出版和管理过程对比

仓储类别	公共数据出版平台	国际数据共享组织或数据中心（WDS）	国家级地学研究机构的数据汇交共享平台
名称	Pangaea	NISDC 冰雪数据中心	英国国家地学数据中心
国家	德国	美国	英国
主题	以地学为主的综合性学科数据	冰川冻土研究数据	地质、地球物理、地球化学数据
数据来源	①世界各国科研人员提交的数据； ②地学科研项目的数据汇交	① NASA 和 NOAA 所监测或采集的数据； ②美国自然科学基金 (NSF) 资助项目的数据； ③其他与冻土研究相关的科研数据	英国自然环境委员会资助的项目所产生的各类数据
数据提交	①采用 Pangaea ticket system 提交数据； ②由作者将数据编辑成数据表进行提交； ③由数据管理员根据数据参数定义进行检查保存到相应的数据库中	①根据资助方提供不同的提交入口，利用网页和 FTP 服务提交数据； ②提交初步数据描述信息后，由数据编辑进行审核判断是否出版； ③同意出版后要求提交详细的数据文档和元数据，并通过 FTP 服务上传数据	①采用英国国家地球科学数据中心提供的统一数据采集入口； ②根据项目规定，按照元数据规范编辑元数据并提交数据集； ③经数据管理员审核后为提交的数据提供 DOI 注册服务
PID 注册标准	DOI	DOI	DOI
长期保存格式	数据表首选格式是制表符分隔文本文件（ASCII），或 Excel 格式，其他格式的文件要求以 ZIP 格式提交	支持各类数据格式	对专业数据格式进行了规范，其中岩土和环境数据为 *.ags 格式，地球物理数据为 *.las、*.sgy、*.xtf 格式
数据质量控制	数据参数的检查 元数据检查 数据的完整性和一致性检查	元数据检查 数据描述文档 数据集检查（由项目的内部专家进行数据的评议）	数据格式 数据集检查 元数据检查
元数据标准	Pangaea 自定义的元数据	采用 NASA 的 DIF 标准 [18]	参照 ISO19115 地理信息元数据的标准自定义的元数据标准
数据引用规范	参照 DataCite 的标准提供数据的引用	自定义数据引用规范	自定义数据引用规范
开放获取程度	开放	部分开放，部分限制	根据项目的要求进行不同级别的开放
数据发现工具	具有推荐功能的统一数据发现工具；高级查询下载工具 PANGAEA Data Warehouse	统一数据发现工具；集成了 NASA Earthdata 和 IceBridge 的数据检索入口	统一数据发现工具
数据互操作服务	SOAPREST	FTP	WMS、WFS、INSPIRE Service

4 数据与论文的联合出版

以 Nature、Science 为代表的顶级学术期刊开始正式提出出版与论文相关数据的要求，并制定了相关的数据存缴和出版政策。但是真正将数据和学术论文联合进行出版的实践较少，其中比较多的是出版领域和公共数据仓储的深度

知识管理论坛，2019（3）:152−162
DOI：10.13266/j.issn.2095−5472.2019.016

合作，如 Elsevier 的 Science Direct 和 Scopus 均与 Pangaea 仓储进行了联合，在提交学术论文的过程中要求作者将数据存储在 Pangaea 仓储中出版并获得一个可访问的链接地址，才能继续进入到学术论文的出版流程，数据和论文出版之间具有一定程度的制约。Elsevier 平台运用 Pangaea 的关联数据工具，可以直接获得出版的数据。在 Pangaea 的数据平台，数据出版后也提供了引用此数据的 Elsevier 出版论文的 DOI，形成互联。联合出版对于论文质量的控制、数据重用、科学结论的验证都具有重要作用，可提供学术论文和科学数据双向透明化访问。Y. Gil 等[19]对于开放科学环境下的未来地球科学论文的出版方式进行了探讨，提出了未来科学论文将包括数据、软件和可复制出版物多种形式，同时具备在开放科学和数字学术环境下的理想特征：①在公共存储库共享数据、软件和其他研究产品；②可使用开放许可；③元数据可用于描述数据、软件和其他研究产品；④数据、软件和其他研究产品都具有唯一的可持续标识符；⑤可在文章中引用上面所有提及的数字资源。目前，尚未真正实现 Y. Gil 等人所提到的完全整合了所有形式学术成果的融合性出版，但是这种多形态的学术成果（论文、数据、软件、其他数据产品等）的出版和开放获取已经成为未来出版领域的趋势。

5 数据出版关键问题探析

5.1 3 种出版模式在数据共享中的作用

上述 3 种出版模式在地学数据共享过程中所起到的作用有一定的差异，数据期刊的出版方式在学术传播方面具有优势，数据论文的影响力较其他出版方式高，但是数据的获取则需要通过数据仓储实现，这些仓储均为与数据期刊合作或由期刊建议提交数据的仓储，以这种方式出版的数据大多数是经过挑选、处理、计算得到的数据集或者是数据产品，研究人员最关注的不是数据论文而是数据本身。通过仓储出版的数据能够直接融入研究数据的管理过程，缩短了数据发布的周期，有利于数据的获取和重用，国内外数据政策的出台积极推动了开放科学和数据监管，研究数据的管理必须要通过数据仓储来实现，此外大规模基础数据集（如遥感、基础地理等数据）也可通过仓储平台的 FTP 服务或接口服务实现互操作。但是，大多数仓储没有学术性的同行评议，数据质量控制大多数属于技术性检查与仓储认证。联合出版集合了上述两种出版的优势，对于验证科学结论、建立完整的学术研究轨迹非常重要，然而这需要出版商、图书馆或数据监管部门、科研机构等协同建立知识生态链。

5.2 数据出版的同行评议问题

同行评议是数据质量控制的关键，对于数据出版来说同行评议虽然不是必须的，但是这种方式却是增加数据可信度的黄金标准。然而目前在数据出版中缺少针对科学数据的同行评议体系。出版领域对数据论文大多数采用了与学术论文相似的同行评议方式来控制数据质量，这种方式的优势在于利用原有的学术刊物的影响力带动了研究人员对数据论文可信度的认可。但是传统的学术论文和数据论文所关注的重点不同，数据出版更注重数据重用这一特点，而学术论文更加注重科学发现的创新性。B. Lawrence[20]等曾经提出科学数据的同行评议通常从数据质量、元数据质量、其他通用因素 3 个方面开展评议。ESSD 采用的两段式的开放式同行评议，利用互联网开放周期让真正的数据使用者评判其数据集的质量，数据用户的反馈对于验证数据和论文内容的一致性和数据质量具有重要意义。此外数据评议的时间选择也是非常关键的，屈宝强等提出随着数据量以指数规模增长，可能更多会选择出版后的同行评议[21]。出版后的评议形式包括针对出版数据的意见征集、数据使用度量和数据修改，可能对数据使用人员而言，更具可扩展性。此外一些基于仓储的数据出版也引入了同行评议的理念和方法，如美国冰雪数据中心，利用项目内部专家对提交到该中心的数据进行评议，由于评

审专家具备相同的专业知识，了解数据重用的方法，对于数据质量的控制具有一定的权威性。数据的评审应该有别于传统的学术论文，不能仅从数据论文的学术性的角度来筛选，而更应该重视数据在参与科研和产生再生性成果的过程中的重用性、元数据的质量、数据使用描述是否完整全面等问题。数据同行评议的专家需要具有相同的专业知识背景和使用同类数据的经验。

5.3 地学数据出版中分层元数据的重要性

元数据主要用于描述数据，帮助研究者实现数据重用，笔者调研的数据质量控制都包含对元数据的检查。在数据的实际应用中，元数据是需要分层次描述的。首先，对于数据使用需求可分为数据发现、数据引用、专业数据描述 3 个层次。发现层通常采用DC核心元数据标准，引用层主要采用或参考 DataCite 的元数据标准，而专业数据描述较为复杂，地球科学领域元数据区别于通用元数据的最显著特点是其数据本身具有的时空特性，地学领域常用的元数据标准有 NASA 的 DIF、ISO19115、ISO19139、FGDC 等，大多数据仓储在专业元数据的描述上都采用或参考了以上元数据标准，有的甚至还同时提供了多种标准的元数据。其次，地学数据的体量较大，基于数据组织的需求，应该根据数据集颗粒度的大小提供多层次的元数据。笔者所调研的数据仓储中，多数规定了单个数据集的大小不超过 1-2G，而对于原始采集的数据来说可能远大于这个体量。如地震反射剖面数据是按照剖面上的接受器进行组织的，遥感数据按照地球的经纬度进行网格化组织，整个研究的地理范围可能包含了若干个数据集。Pangaea 在出版大体量的反射地震数据时，将数据集拆解成若干个数据序列，赋予每个数据序列一个唯一的标识符，这时既需要给每一个数据序列提供专业元数据，同时也需要提供整个研究区域数据集的元数据信息。综上所述在地学数据出版实践中，元数据的分层描述对于数据的保存和重用都具有重要的意义。

6 总结与展望

从以上研究可以发现，现有地学数据期刊的载文量呈现逐年上升的趋势，说明数据出版逐渐被科研人员认可与接受，这对于研究者积极地共享数据具有促进作用。国内外科学数据管理政策的出台使得科学数据的保存与管理成为研究中必不可少的环节，而数据出版与数据管理息息相关，任何模式的数据出版都离不开数据仓储。笔者提取了数据出版必要的 5 个要素，但是通过调研发现注册在 re3data 中的地学仓储能够完全满足这 5 要素的为数不多，这意味着大多数数据仓储尚未具备完善的数据出版能力。学术论文和数据的联合出版受到开放科学环境的限制，目前实践较少。综合调研的结果，笔者对其存在的关键问题进行了探讨，这对于我国地学领域中数据共享模式向数据出版转变的实践具有借鉴意义。

此外，地学数据出版具有自身的领域特点，地学研究的过程通常被概括为 3 个阶段：通过仪器采集数据，对数据进行分析和处理，通过创建研究方法生成数据产品和研究结论。在这 3 个阶段中可能产生的数据包含：原始采集的数据，数据读取、转换、可视化等软件，由于创建新的数据处理或计算方法而产生的新的数据产品，融合多种数据及数据衍生物的出版对于地学领域的数据共享将是一个挑战。

参考文献：

[1] 诸云强 , 朱琦 , 冯卓 , 等 . 科学大数据开放共享机制研究及其对环境信息共享的启示 [J]. 中国环境管理 , 2015, 7(6):38-45.

[2] 王卷乐 , 孙九林 . 世界数据中心 (WDC) 回顾、变革与展望 [J]. 地球科学进展 , 2009, 24(6):612-620.

[3] 诸云强 , 孙九林 , 廖顺宝 , 等 . 地球系统科学数据共享研究与实践 [J]. 地球信息科学学报 , 2010, 2010(1):1-8.

[4] 吴立宗 , 南卓铜 , 王亮绪 . 科学数据出版——促进数据共享的一种新模式 [J]. 中国科技资源导刊 , 2014(5):72-78.

[5] EAR Division Data Sharing Policy [EB/OL]. [2018-12-18]. https://www.nsf.gov/geo/geo-data-policies/ear/ear-da-

知识管理论坛，2019(3):152–162
DOI：10.13266/j.issn.2095–5472.2019.016

ta-policy-apr2018.pdf.

[6] 何琳 , 常颖聪 . 国内外科学数据出版研究进展 [J]. 图书情报工作 , 2014, 58(5):104-110.

[7] BRASE J, FARQUHAR A, GRUTTEMEIER H, et al. Approach for a joint global registration agency for research data[J]. Information services & use, 2009, 29(1):13-27.

[8] KRATZ J E, STRASSER C. Researcher perspectives on publication and peer review of data[J]. PLOS ONE, 2015, 10(2):e0117619.

[9] 李丹丹 , 吴振新 . 研究数据引用研究 [J]. 图书馆杂志 , 2013, 32(5):65-71.

[10] PRONK T E，WIERSMA P H，VAN WEERDEN A，et al. A game theoretic analysis of research data sharing［J］. Peer J，2015（3）：e1242.

[11] WDS[EB/OL].[2018-12-18].http://www.icsu-wds.org/organization/intro-to-wds.

[12] WATERMAN K J, SIERMAN B. Survey of DSA-certified digital repositories : report on the findings in a survey of all DSA-certified digital repositories on investments in and benefits of acquiring the Data Seal of Approval (DSA) [R]. Hague: Netherlands Coalition for Digital Preservation, 2016:19.

[13] CoreTrustSeal[EB/OL].[2019-01-12]. https://www.coretrustseal.org/about/.

[14] 黄如花 , 李楠 . 开放数据的许可协议类型研究 [J]. 图书馆 , 2016(8):16-21.

[15] Data Citation Synthesis Group. Joint Declaration of Data Citation Principles [EB/OL]. [2018-02-18].https://www.force11.org/group/joint-declaration-data-citation-principles-final.

[16] DataCite Metadata Working Group. DataCite Metadata Schema for the Publication and Citation of Research Data[EB/ OL].[2018-02-18]. http://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf.

[17] 王丹丹 . 科学数据出版平台的用户测试研究 [J]. 情报资料工作 , 2017(6):58-63.

[18] National Aeronautics and Space Administration—Global Change Master Directory. Directory Interchange Format (DIF) Writer's Guide[EB/OL].[2017-10-15]. http://gcmd.nasa.gov/add/difguide/.

[19] GIL Y, DAVID C H, DEMIR I, et al. Toward the geoscience paper of the future: best practices for documenting and sharing research from data to software to provenance[J]. Earth and space science, 2016, 3(10): 388-415.

[20] LAWRENCE B, JONES C, MATTHEWS B, et al. Citation and peer review of data: moving towards formal data publication[J]. International journal of digital curation, 2011, 6(2):4-37.

[21] 屈宝强 , 王凯 . 数据出版视角下的科学数据同行评议 [J]. 图书馆杂志 , 2017, 36(10):71-77.

作者贡献说明：

韩　露：论文相关的数据采集，数据出版关键问题的研究与论文撰写；

丁　毅：数据仓储研究和数据出版流程实践。



知识管理论坛
2019 年 第 3 期(总第 21 期)

A Contrastive Study of Practical Modes of Data Publishing

——Take the Field of Earth Science as an Example

Han Lu[1] DingYi[2]

[1] Beijing Institute of Technology Library, Beijing 100081

[2] Institute of Geology, Chinese Academy of Geological Sciences, Beijing 100037

Abstract: [Purpose/significance] Scientific data publishing is the main mode of academic communication for data-intensive scientific discovery, which is of great significance for data reuse and scientific verification. **[Method/process]** Earth sciences has undergone a great change from the data sharing model to the current data publishing. Current practices of data publishing can be divided into three modes: data journal publishing, data warehousing publishing, data and paper joint publishing. The author made statistics and comparison on the practice methods and key elements of each mode, and emphatically analyzed the advantages and disadvantages of the three modes, peer data review issues and the importance of hierarchical metadata in geoscience data publishing. **[Result/conclusion]** Through research, the author proposes that warehouse-based publishing facilitates integration into scientific data management process and facilitates data reuse. However, such publishing mode usually lacks peer review. Peer review of data should be different from academic papers and should focus on the reusability of data in the process of participating in scientific research and producing regenerative results. The hierarchical description of metadata is of great significance to the preservation and reuse of geoscience big data.

Keywords: data publication data repository data journal geoscience data